

Statistics and Data Science 171  
YData: Text Data Science

# Overview of Topic Models

February 21



# Intro to Topic Modeling

Some of the following slides are from Dave Blei's tutorial on Topic Modeling

<http://www.cs.columbia.edu/~blei/topicmodeling.html>

A survey paper describing many of these ideas in more detail is here:

[https://cacm.acm.org/magazines/2012/4/  
147361-probabilistic-topic-models/fulltext](https://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext)

# Topic modeling



Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

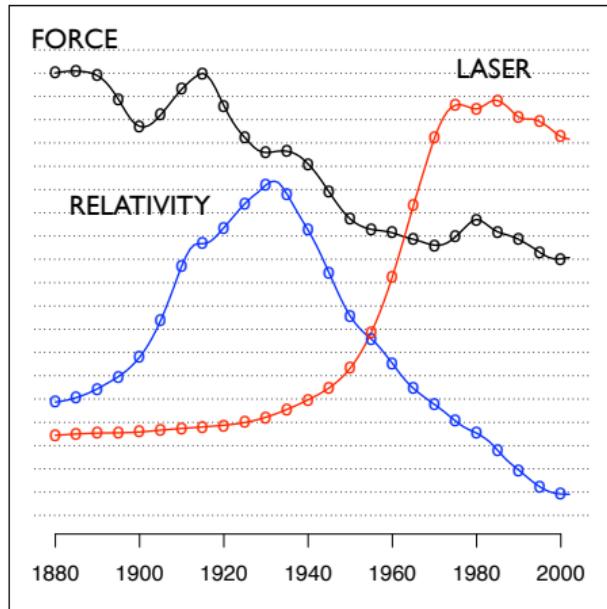
- ① Discover the hidden themes that pervade the collection.
- ② Annotate the documents according to those themes.
- ③ Use annotations to organize, summarize, and search the texts.

# Discover topics from a corpus

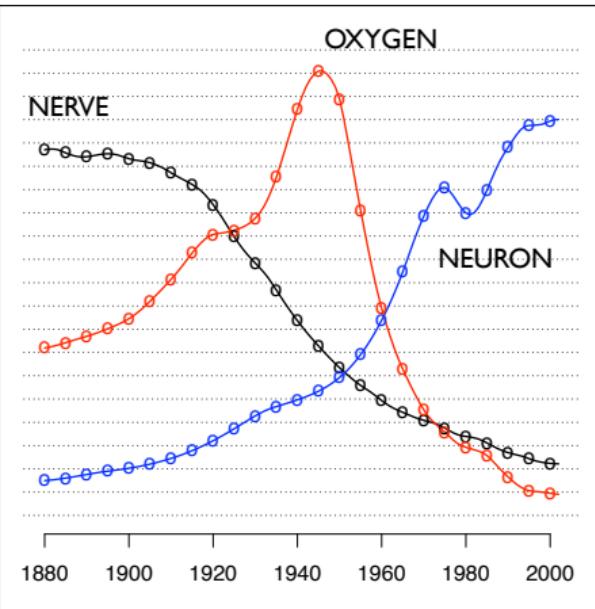
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

# Model the evolution of topics over time

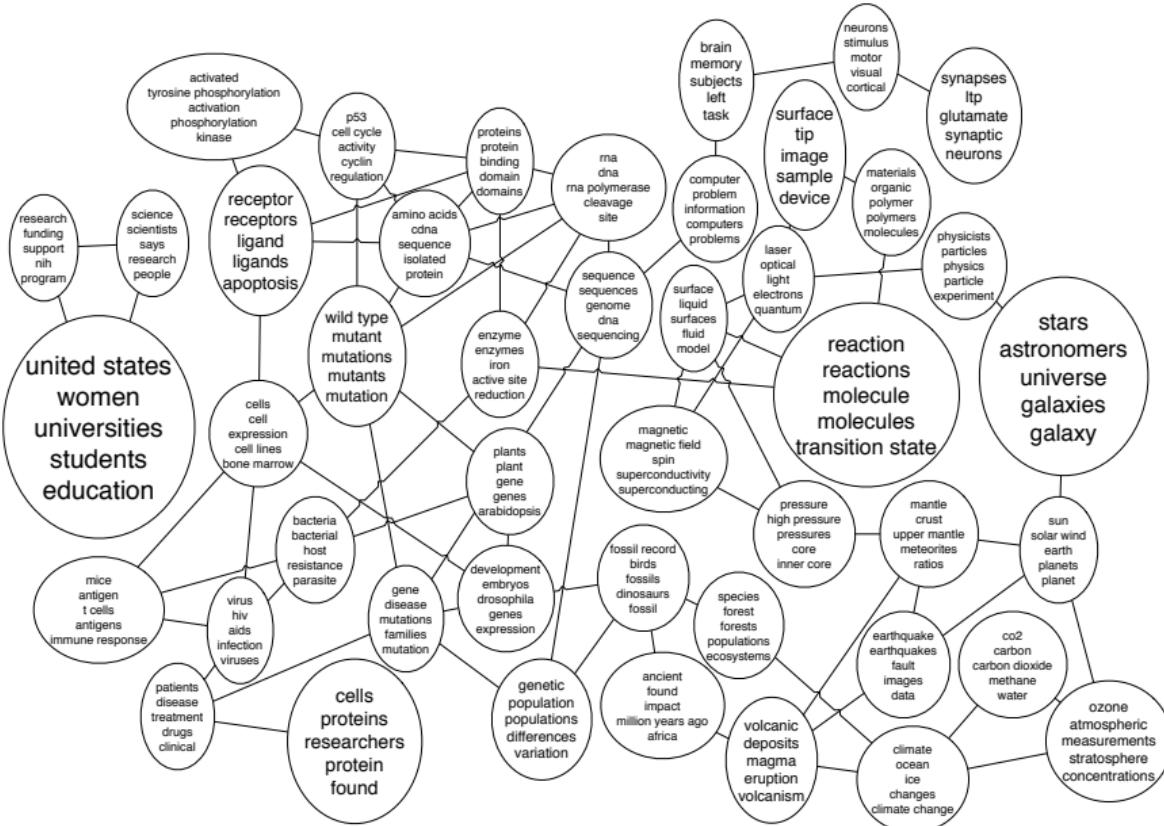
"Theoretical Physics"



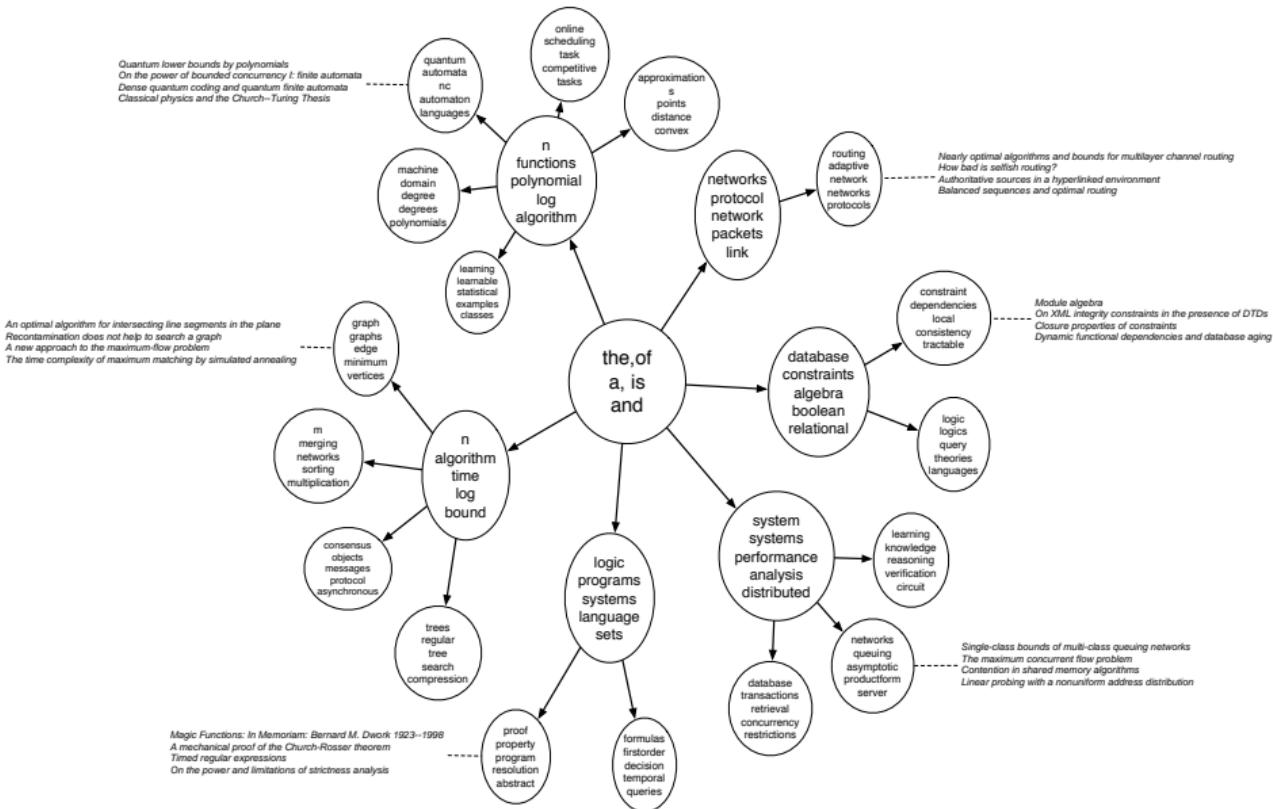
"Neuroscience"



# Model connections between topics



# Find hierarchies of topics



# Annotate images



SKY WATER TREE  
MOUNTAIN PEOPLE



SCOTLAND WATER  
FLOWER HILLS TREE



SKY WATER BUILDING  
PEOPLE WATER



FISH WATER OCEAN  
TREE CORAL



PEOPLE MARKET PATTERN  
TEXTILE DISPLAY



BIRDS NEST TREE  
BRANCH LEAVES

# Captioning

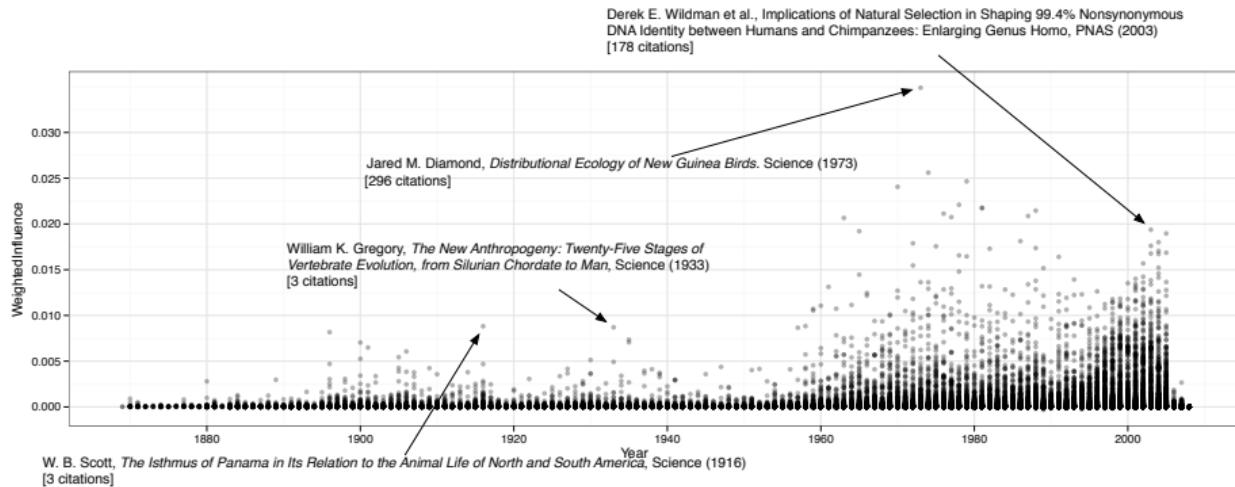


there is a large bird on the water  
a small bird sitting on top of a lake  
a large white bird standing on the water on a beach  
a bird is on the water on a beach  
a bird that is standing in the water



a professional baseball game is played in the middle of the field  
several players at the end of a baseball game  
a group of players playing a baseball game  
the baseball players are playing games at the field  
a baseball players are playing with a game and fans

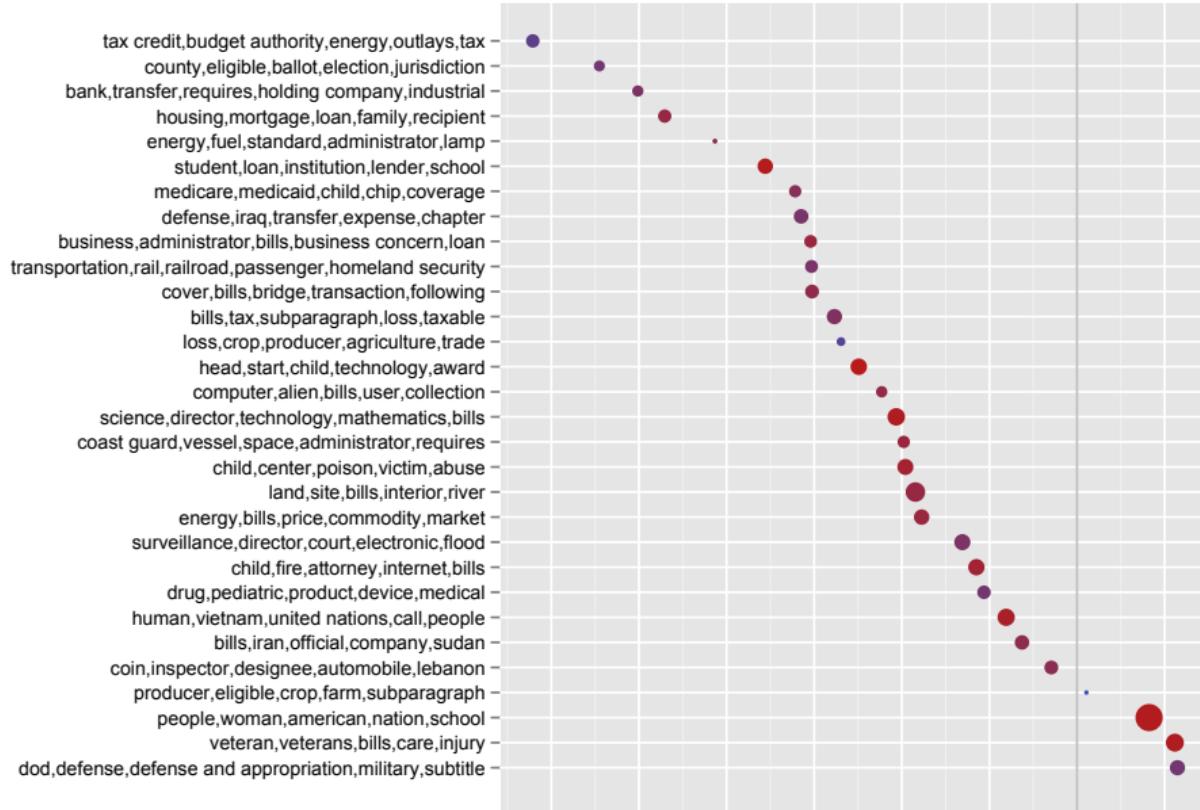
# Discover influential articles



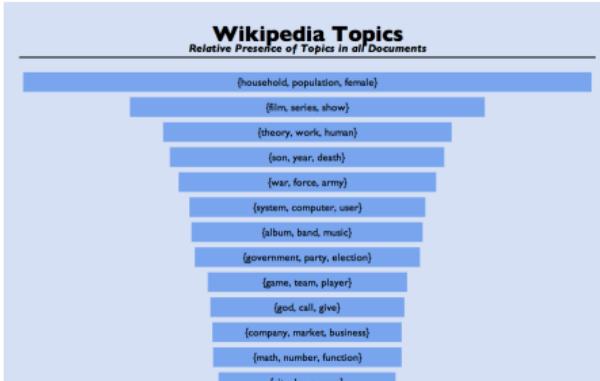
# Predict links between articles

<p><i>Markov chain Monte Carlo convergence diagnostics: A comparative review</i></p> <p><b>Minorization conditions and convergence rates for Markov chain Monte Carlo</b></p> <p>Rates of convergence of the Hastings and Metropolis algorithms</p> <p><b>Possible biases induced by MCMC convergence diagnostics</b></p> <p>Bounding convergence time of the Gibbs sampler in Bayesian image restoration</p> <p>Self regenerative Markov chain Monte Carlo</p> <p>Auxiliary variable methods for Markov chain Monte Carlo with applications</p> <p><b>Rate of Convergence of the Gibbs Sampler by Gaussian Approximation</b></p> <p>Diagnosing convergence of Markov chain Monte Carlo algorithms</p>	<b>R&amp;TM (<math>\psi_e</math>)</b>
<p>Exact Bound for the Convergence of Metropolis Chains</p> <p>Self regenerative Markov chain Monte Carlo</p> <p><b>Minorization conditions and convergence rates for Markov chain Monte Carlo</b></p> <p>Gibbs-markov models</p> <p>Auxiliary variable methods for Markov chain Monte Carlo with applications</p> <p>Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models</p> <p>Mediating instrumental variables</p> <p>A qualitative framework for probabilistic inference</p> <p>Adaptation for Self Regenerative MCMC</p>	<b>LDA + Regression</b>

# Characterize political decisions



# Organize and browse large corpora



**{film, series, show}**

words	related documents	related topics
film	The X-Files	{son, year, death}
series	Orson Welles	{work, book, publish}
show	Stanley Kubrick	{album, band, music}
character	B movie	{woman, child, man}
play	Mystery Science Theater 3000	{law, state, case}
make	Monty Python	{black, white, people}
episode	Doctor Who	{theory, work, human}
movie	Sam Peckinpah	{@card@, make, design}
good	Married... with Children	{war, force, army}
release	History of film	{god, call, give}
feature	The A-Team	{game, team, player}
television	Pulp Fiction (film)	{day, year, event}
star	Mad (magazine)	{company, market, business}

**Stanley Kubrick**

Topic	Percentage
{film, series, show}	26%
{theory, work, human}	15%
{son, year, death}	12%
{war, force, army}	10%
{god, call, give}	8%
{math, number, function}	5%
{company, market, business}	4%
{black, white, people}	3%
{theory, work, human}	2%
{theory, work, human}	1%

**related topics**

- {film, series, show}
- {theory, work, human}
- {son, year, death}
- {war, force, army}
- {god, call, give}
- {math, number, function}
- {company, market, business}
- {black, white, people}
- {theory, work, human}
- {theory, work, human}

**related documents**

- Orson Welles
- B movie
- Mystery Science Theater 3000
- Monty Python
- Doctor Who
- Sam Peckinpah
- The A-Team
- Pulp Fiction (film)
- Buffy the Vampire Slayer (TV series)
- The X-Files
- Sunset Boulevard (film)
- Jack Benny

**{theory, work, human}**

words	related documents	related topics
theory	Meme	{work, book, publish}
work	Intelligent design	{law, state, case}
human	Immanuel Kant	{son, year, death}
idea	Philosophy of mathematics	{woman, child, man}
term	History of science	{god, call, give}
study	Free will	{black, white, people}
view	Truth	{film, series, show}
science	Psychoanalysis	{war, force, army}
concept	Charles Peirce	{language, word, form}
form	Existentialism	{@card@, make, design}
world	Deconstruction	{church, century, christian}
argue	Social sciences	{rate, high, increase}
social	Idealism	{company, market, business}

# Topics in scientific texts

<b>Quantum physics</b>	spin energy field electron magnetic state states hamiltonian
<b>Particle physics</b>	higgs neutrino coupling decay scale masses mixing quark
<b>Astrophysics</b>	mass gas star stellar galaxies disk halo radius luminosity
<b>Relativity</b>	black metric hole schwarzschild gravity holes einstein
<b>Number theory</b>	prime integer numbers conjecture integers degree modulo
<b>Graph theory</b>	graph vertex vertices edges node edge number set tree
<b>Linear algebra</b>	matrix matrices vector basis vectors diagonal rank linear
<b>Optimization</b>	problem optimization algorithm function solution gradient
<b>Probability</b>	random probability distribution process measure time
<b>Machine learning</b>	layer word image feature sentence model cnn lstm training

# Topics modeling for equations

Topic	Generated Equations
Quantum physics	<ul style="list-style-type: none"><li>• <math>E = \hbar \frac{\partial^2 S}{\partial t^2} \left( \frac{\partial \varphi}{\partial c} \right) - \frac{k}{\hbar^2} \frac{\partial B}{\partial t} (t + \partial_t \delta).</math></li><li>• <math>\Psi_{\text{pr}} = \sum_{\mathbf{l}} (\psi_{\mathbf{r}+\uparrow} - \psi_{\mathbf{r}\downarrow}^\dagger) + \sum_{\mathbf{r}'} (\psi_{\mathbf{r}\downarrow,\uparrow}^\dagger - \psi_{\mathbf{r}\downarrow} \sigma^\dagger).</math></li></ul>
Particle physics	<ul style="list-style-type: none"><li>• <math>\mathcal{H} = \frac{1}{4}(\partial_\mu \phi)^2 + 2m\phi_\nu(\phi) + \frac{1}{2}m^2(\phi)(1 - \phi^2)^2.</math></li><li>• <math>m_{\text{eff}}(M) = 1.4 \cdot 10^{-13} \text{ GeV}.</math></li></ul>
Relativity	<ul style="list-style-type: none"><li>• <math>\mathcal{M} = \frac{1}{2}g^{\mu\nu}(f_{\mu\nu,\mu} - g_{\mu\nu,\nu} + g_{\nu\nu,b}f_{\mu,\nu}) + \frac{1}{2}g^{\mu\nu}.</math></li><li>• <math>T_{\mu\nu} = \int_0^\infty ds_{\mu\nu} ds^2 + a_\mu^2 dr^2 + r^2 d\Omega^2.</math></li></ul>

Michihiro Yasunaga:

[https://michiyasunaga.github.io/files/papers/TopicEq\\_aaai19\\_final.pdf](https://michiyasunaga.github.io/files/papers/TopicEq_aaai19_final.pdf)

<https://seas.yale.edu/news-events/news/michihiro-yasunaga-wins-cra-outstanding-undergraduate-researcher-award>

# Uber Topics

From a former student:

*I just wanted to let you know that I'm currently using Latent Dirichlet Allocation at my current job at Uber!*

*We're using LDA to discover topics in rider feedback – when riders write comments about their driver after the trip. We're trying to find topics such as 'unprofessional driver', 'driver no-show', 'sexual harassment', etc. LDA has worked really well with this – so thank you for covering it in much detail in your course.*

# **Introduction to Topic Modeling**

# Probabilistic modeling

- ① Data are assumed to be observed from a generative probabilistic process that includes hidden variables.
  - *In text, the hidden variables are the thematic structure.*
- ② Infer the hidden structure using posterior inference
  - *What are the topics that describe this collection?*
- ③ Situate new data into the estimated model.
  - *How does a new document fit into the topic structure?*

# Latent Dirichlet allocation (LDA)

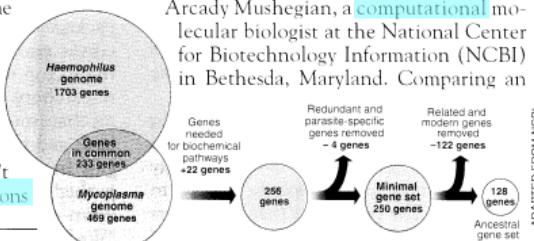
## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

“are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

# Generative model for LDA

Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

Documents

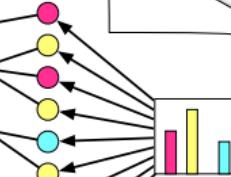
## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>3</sup> two genomic researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Stu Atkinson of Icahn School of Medicine at Mount Sinai University in New York, who arrived at the 800 number. But coming with a consensus answer may be more than just a genetic numbers game, particularly if more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

Topic proportions and assignments



\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

# The posterior distribution

Topics



Documents

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>\*</sup> two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

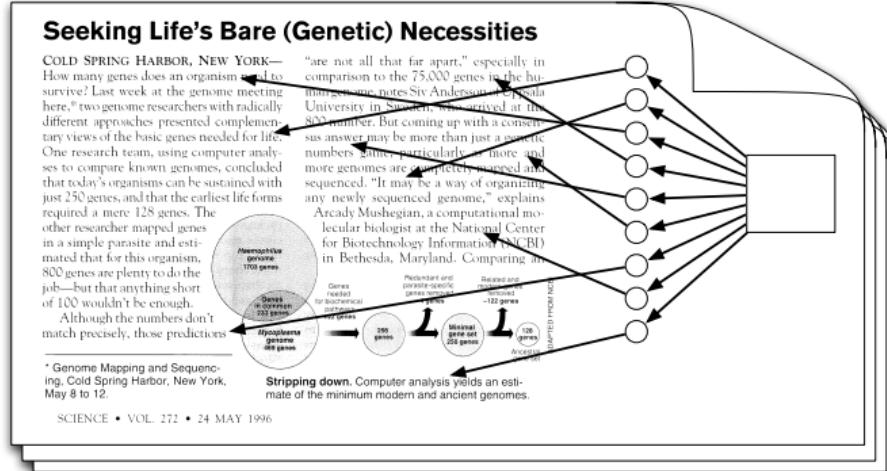
Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Umeå University in Sweden. She arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game; particularly, as more and more genomes are completely sequenced and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

\* Genome Mapping and Sequencing. Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 271 • 24 MAY 1996

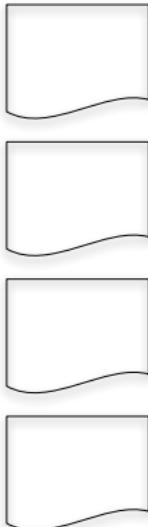
Topic proportions and assignments



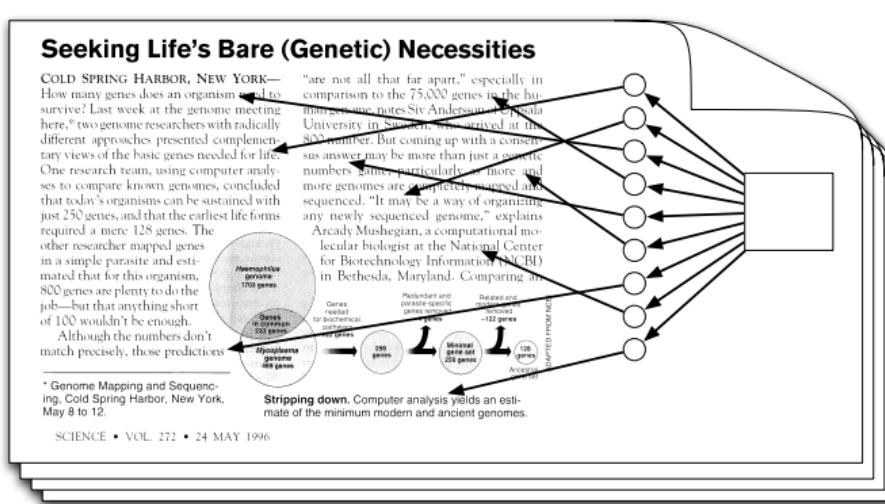
- In reality, we only observe the documents
- The other structure are **hidden variables**

# The posterior distribution

Topics



Documents



Topic proportions and assignments

- Our goal is to **infer** the hidden variables
  - I.e., compute their distribution conditioned on the documents
- $$p(\text{topics, proportions, assignments} \mid \text{documents})$$

# Summary

- Topic models automatically extract “semantic themes” from large document collections
- Based on latent variable models
- Can be useful for a wide variety of data